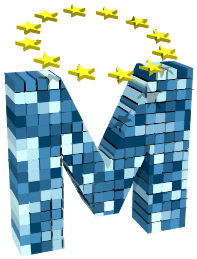


Project title: Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and OWNer control
Project acronym: MOSAICrOWN
Funding scheme: H2020-ICT-2018-2
Topic: ICT-13-2018-2019
Project duration: January 2019 – December 2021

D2.3

Final report on research alignment

Editors: Flora Giusto (MC)
Saverio Mucci (MC)
Reviewers: Stefano Paraboschi (UNIBG)
Pierangela Samarati (UNIMI)

Abstract

This documet details the monitoring of the research and technological work done in WPs3-5, with the objective to ensure the alignment with the Use Cases requirements and development described in D2.1 and D2.2. Each industrial partner provides a final status report for its own use case.

Type	Identifier	Dissemination	Date
Deliverable	D2.3	Public	2021.06.30

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825333.

MOSAICrOWN Consortium

1.	Università degli Studi di Milano	UNIMI	Italy
2.	EMC Information Systems International	EISI	Ireland
3.	Mastercard Europe	MC	Belgium
4.	SAP SE	SAP SE	Germany
5.	Università degli Studi di Bergamo	UNIBG	Italy
6.	GEIE ERCIM (Host of the W3C)	W3C	France

Versions

Version	Date	Description
0.1	2021.06.04	Initial Release
0.2	2021.06.25	Second Release
1.0	2021.06.30	Final Release

List of Contributors

This document contains contributions from different MOSAICrOWN partners. Contributors for the chapters of this deliverable are presented in the following table.

Chapter	Author(s)
Executive Summary	Flora Giusto (MC), Saverio Mucci (MC)
Chapter 1 : Introduction	Flora Giusto (MC), Saverio Mucci (MC)
Chapter 2 : Use Case 1 (EISI)	Aidan O Mahony (EISI)
Chapter 3 : Use Case 2 (MC)	Flora Giusto (MC), Saverio Mucci (MC)
Chapter 4 : Use Case 3 (SAP SE)	Jonas Böhler (SAP SE)
Chapter 5 : Conclusions	Flora Giusto (MC), Saverio Mucci (MC)

Contents

List of Figures

List of Tables

Executive Summary

Data protection is key for public and private organizations who must design effective data strategies meeting regulations, like GDPR in Europe, consumers expectations and nurture their continuous innovations development. Indeed, privacy regulations are creating both opportunities and expectations. Innovation is enabling organizations to create a differentiation advantage and build trust with their consumers. The necessity to build privacy-by-design as a standard business process is more and more important due to the regulation requirements. On the other hand, the new market technologies enable privacy-compliant analytics thanks to self-service tools that have made analytics accessible without data science expertise, mainly via cloud-based platforms, which also contributed to reduce entry barriers. In this context, MOISAICrOWN, a consortium compromising academic and business partners, is cooperating to develop solutions that are supporting both data disclosure and data sanitization. This deliverable provides a final analysis on research alignment covering the activity of Task 2.2. Task 2.2 is part of Work Package 2 that oversees the Use Cases requirements, deployment and validation with a continuous monitoring. Deliverable D2.3 builds on what was presented in D2.2 and showcases the three use cases which provide their own scenarios for MOSAICrOWN's data protection platform. D2.3 details the monitoring of the research and technological work done in the WPs3-5, with the objective to ensure the alignment with the use cases requirements and development described in D2.1 and D2.2. The academic and the industrial partners have used the methodology described in the introduction to carry out the monitoring. The results for each use case are presented in a dedicated chapter with emphasis on the key features of the targeted Use Case. An analysis of the whole of the results is detailed at the end of each use case chapter and shows how the use case contributes to the key dimensions of data protection. Use Case 1, defined by EISI, focuses on Intelligent Connected Vehicles (ICV) and, more specifically, how sensitive data is passed from a charging infrastructure to the data market. Use Case 2, led by Mastercard, focuses on transaction-level financial data and the importance of data wrapping techniques. Use Case 3, overseen by SAP SE, takes on a broader set of operational and experience data used for consumer analytics while considering a cloud-based scenario.

1. Introduction

The goal of MOSAICrOWN is to enable data sharing and collaborative analytics in multi-owner scenarios, ensuring proper protection of private/sensitive/confidential information. MOSAICrOWN provides effective and deployable solutions allowing data owners to maintain control on the data sharing process, enabling selective and sanitized disclosure providing for efficient and scalable privacy-aware collaborative computations. This document is the final report on the alignment between the research and technological development in WPs3-5 and the use cases.

1.1 Context and purpose of this document

The goal of Work Package 2 (WP2) is to coordinate the use cases considered in the project, provide requirements, deployment, and validation of MOSAICrOWN solutions, to enable direct exploitation by the industrial partners.

For the original list of requirements, along with a summary of previous progress, please see D2.2. This document builds on D2.2 and provides an update on each individual use case as a final report.

Figure 1.1 visualizes the interaction between the different work packages in MOSAICrOWN, which may be useful to review before continuing with the document.

1.2 Continuous Monitoring

1.2.1 Work package summary

Before moving into the following chapters, it is important to provide a brief summary of MOSAICrOWN's work packages, as they are referenced frequently throughout. While these summaries are short, further information can be found regarding the work packages in either D2.1, D2.2 or in the deliverables of WP3-5 themselves.

Work Package 2 coordinates the use cases considered in the project, providing requirements, deployment and validation of MOSAICrOWN solutions, and enabling direct exploitation by the industrial partners.

Work Package 3 is responsible for defining the data governance framework for the management of data in collaborative data platforms. This includes concertation of the variety of available tools, a policy model and accompanying language, enabling owners to comply effectively with the requirements identified in WP2 and to share data with confidence and control on the data value chains.

Work Package 4 is responsible for designing techniques able to efficiently support the protection requirements expressed by the policy produced in WP3. This work leads to the realization of a platform for the protection of personal/sensitive/confidential information in domains involving

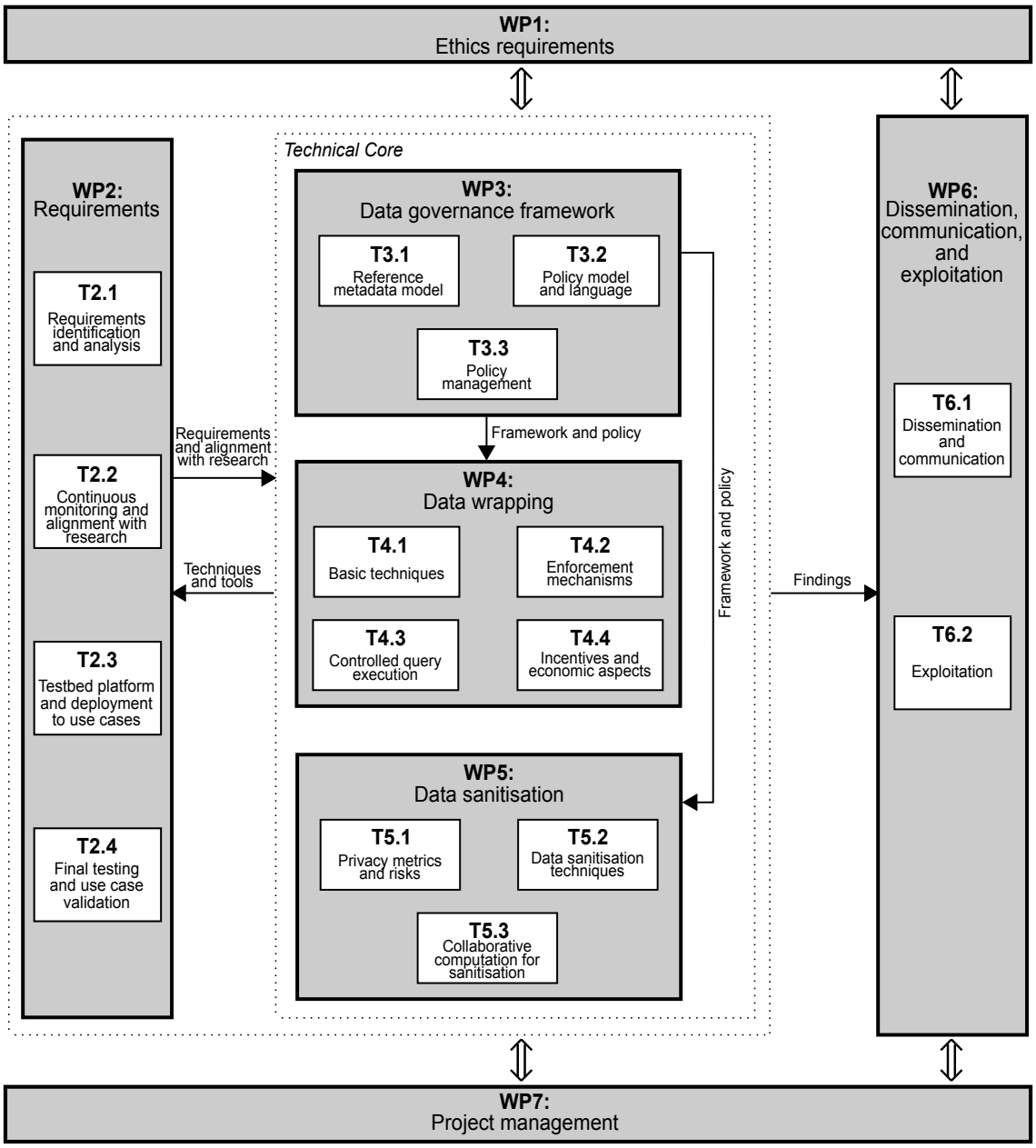


Figure 1.1: MOSAICrOWN Work packages overview

multiple data owners. The developed techniques guarantee processing and sharing of data, adopting a variety of approaches: encryption, hashing, tokenization. Attention is made to guarantee that the manager of the data market is unable to violate the protection policy demanded by the data owners. As an advanced evolution of the scenario, the platform also permits to benefit from the availability of a collection of providers to implement data storage and collaborative query management approaches, accounting for different trust and control of data exhibited by each provider.

Work Package 5 is responsible for developing techniques for protecting personal /sensitive /confidential information in collaborative computations and/or data sharing/release/publication. The techniques developed WP 5, leveraging and extending approaches for data anonymization and obfuscation, produce sanitized versions of the underlying (structured or unstructured) data for

the privacy-preserving use, sharing, and computation with other parties. The WP identifies privacy metrics and enables privacy assessment to better understand the privacy protection enjoyed by the sanitized data. It also focuses on utility assessment and metrics. The techniques are designed for enabling their efficient implementation to ensure performance and therefore actual applicability of the tools developed.

Work Package 6 coordinates and oversees the dissemination, communication, and exploitation activities within the project and organizes collaboration with other research efforts addressing similar goals.

1.2.2 Scope of the monitoring

The scope of the monitoring covers the WPs in which the research and technological work is been undertaken, and the use cases described,as follows.

Targeted WPs. The work packages targeted for monitoring are WP3, WP4 and WP5 dealing with the core technical research and development activities.

Targeted use cases. The targeted use cases for the monitoring (as defined in T2.2) are:

- Use Case 1: Protection of Sensitive Data in an Intelligent Connected Vehicle (ICV)
- Use Case 2: Data sharing and analysis via data anonymization and compliance
- Use Case 3: Cloud-based data market for privacy-preserving consumer analytics

Each use case addresses a specific application scenario. However, taken together, the use cases provide a suite of capabilities that cover data protection as ultimate target of MOSAICrOWN.

1.3 Structure of the document

In this chapter, we outlined the overall purpose of the document and provided an overview of the work packages.

Chapters 2 through 4 cover updates regarding each use case, respectively. Each chapter is structured in the same way: it begins with a technical overview of the use case, monitoring of the tasks, deployment status, analysis of the monitoring and findings corelating the elements just listed with the relevant aspects of WPs 3-5.

The deliverable concludes with a chapter summarizing the general findings and providing a status update on what is to come.

2. Use Case 1 (EISI)

Use Case 1 (UC1) is concerned with securing personal data in the context of an Intelligent Connected Vehicle (ICV). Furthermore, UC1 is interested in the semantic fusion of data from electric vehicles and electric vehicle charging stations such that fleet owners, ICV drivers, and charging infrastructure providers can mutually benefit from data while at the same time ensuring the data is protected.

The protection of personal data in UC1 is based on both application of policy, based on the policy language presented in D3.3 ("First version of policy specification language and model"), and the data wrapping of data, based on the tools presented in D4.1 ("First version of encryption-based protection tools"). The combination of these protections are referred to as data governance.

The goals of the tools provided for UC1 are:

- Allowing analytics on personal data of ICV drivers, the ICV's themselves, and the usage of the EV charging infrastructure
- Providing mechanisms for data ingestion. The sources of the ingestion are the ICV's, policy specifications from drivers, usages of the charging infrastructure, and the data from the users of the governance framework
- Enabling users of the marketplace to gain insights into various aspects of the data being generated as well as facilitating access to that data based on policy enforcement
- Safeguarding personal data through data wrapping techniques

For this deliverable we consider the final research alignment when viewed together along with the deployment of the tools, the satisfaction of the requirements, and the enhancement of the research in the field of data governance.

2.1 Technical overview

Use Case 1 deals with ingestion of sensitive data during the data life-cycle an is concerned with policy application, sanitization, and wrapping as highlighted in Figure 2.1. UC1 facilitates the ingestion of disparate data sources for the purpose of allowing multiple parties develop monetized services. These services exist within the context of ICV data and will allow for the creation of new data markets. There are significant opportunities within the automotive industry for smart, connected technologies facilitating intelligent vehicles as well data markets. Outside of the automotive industry, there are many other industries eager for taking advantage of massive datasets within feature rich data markets. An example of such an industry is that of smart cities and energy conservation. UC1 aligns itself with this industry via the analytics provided by ICV charging stations.

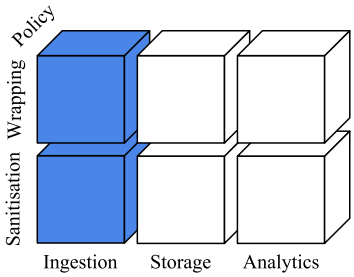


Figure 2.1: Overview of Use Case 1 dimensions

In order to satisfy the requirements of UC1 there is an obvious need for data. The data needed by UC1 drives new monetization opportunities as well as enhancing the customers experience through the exploitation of their own personal data at their own level of comfort. This level of comfort needs to be considered and methods on enforcing the data regulations are relevant to discuss at this point. These regulations are used to provide ethical uses of data, however the risk of hindering innovation is one we need to account for in our tools. The satisfaction of the requirements of UC1 allow for the demonstration of how MOSAICrOWN can both be effective in a real-world situation at data protection as well as facilitate a vibrant ICV data market. UC1 is illustrated in Figure 2.2. The actors involved in the use case are: the connected vehicle fleet, the electric vehicle (EV) charging infrastructure provider, and the MOSAICrOWN cloud provider hosting the data sharing and analytics platform. Source of data include the driver of the EV, the EV itself, the fleet owner, and the EV charging infrastructure. All this data is integrated as part of the data market services.

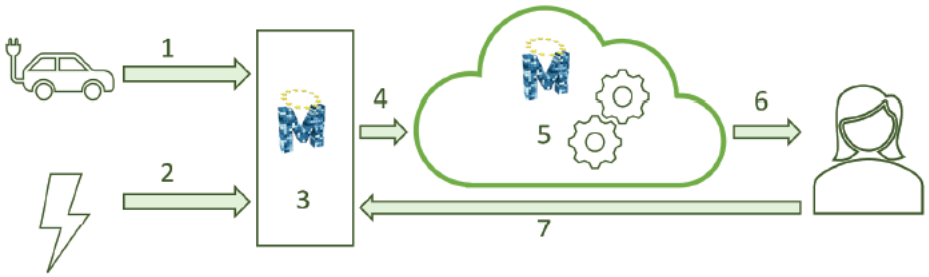


Figure 2.2: UC1 Overview

2.2 Monitoring of the tasks

The following subsection provides an overview of the monitoring and alignment tasks of the technical work packages with regards to UC1.

2.2.1 WP3 monitoring and alignment

WP3 is responsible for the data governance framework, the policy model and language. It aligns closely with WP2 to facilitate the policy-based data access as well as assisting in the privacy preservation aspects of the testbed developed in WP2. UC1 benefited from the policy management task (T3.3) which allowed the identification of policies suitable to enable implementation of UC1 related requirements, e.g., the policy choice available to the EV driver. The testbed deployment

integrated the policy language into the customer web application developed which allowed for converting between JavaScript Object Notation (JSON) to JavaScript Object Notation for Linked Data (JSON-LD) to N-Triples.

The testbed platform is discussed in more detail in Section 2.3 where we present the automotive application which facilitates the EV driver to select what policy they want applied to their data in such a way as to allow for the driver to benefit in a monetary fashion from the use of their personal data. Furthermore, the automotive application allows for the data ingestion from the EV as well as displaying data relevant to the EV and the EV driver as the EV journeys progress. We also present our development of a customer web application to engage with the MOSAICrOWN data market, in terms of analytics as well as accessing raw data (policy permitting). This web application also allows for various roles within the data market, thus best facilitating privacy preservation as well as monetization via role-based analytics.

2.2.2 WP4 monitoring and alignment

WP4 is responsible for the development of tools and techniques designed to protect the requirements expressed by the policy in WP3. UC1 made use of tools developed in T4.2 and T4.3 for the protection of data from the EV (both driver personal data and the actual vehicle data). This integration also demonstrates the integration of the tools developed in MOSAICrOWN and, furthermore, provides for greater performance of the data wrapping functionality of the data governance framework.

2.2.3 WP5 monitoring and alignment

WP5 in relation to UC1 is concerned with the sanitization of data being ingested into the governance framework. EISI is contributed to T5.1 “Privacy metrics and risks” which completed in M24, however the application of these metrics will be integrated into the testbed platform delivered as part of T2.3 “Testbed platform and deployment to Use Cases”.

2.3 Deployment status

The goal of MOSAICrOWN is to apply data analysis techniques for the provision of analytics over large collections of data. Furthermore, it provides these techniques in a manner which ensures proper protection of private or sensitive data. UC1 further develops techniques specific to the automotive world. The three tools developed to enable the deployment of UC1 are presented in this section and are further highlighted in context of the overall platform in Figure 2.3. The three highlighted components in this platform diagram are: the automotive application for installation within an EV which facilitates the EV data ingestion, the customer web application for customer interaction with the MOSAICrOWN data market, and the data wrapping tools required for the protection of sensitive data within the data market.

2.3.1 Automotive policy selection and data ingestion

The tool developed to address automotive ingestion to the data market is described in this subsection. It addresses a number of requirements presented in the Deliverable D2.1 (“Requirements from the Use Cases”). Based on an evaluation of the available IVI systems. We decided to use Android Auto to develop the MOSAICrOWN mobile application. we decided to allow the EV

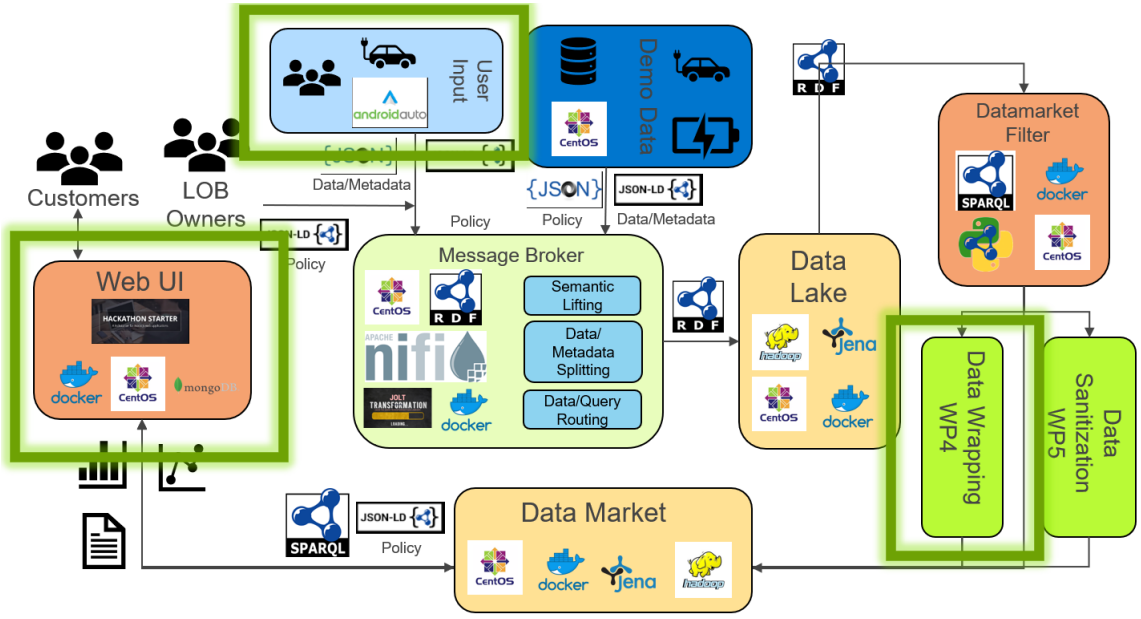


Figure 2.3: UC1 Highlighted components developed for deployment

driver to decide on what policy is applied to their data. To simplify the interaction for the EV driver we predefined three levels of policies. These are (a) a policy with the most private settings described, (b) a policy with "moderate" privacy settings stipulated, and (c) a policy with the least privacy preserving settings described. Figure 2.4 illustrates how this is presented to the EV driver.

2.3.2 Customer web application

The main functionality of web tools is to enable data access by different users (MOSAICrOWN Cloud Provider, Car Driver, Fleet Owner and EV Charging Infrastructure Provider) according to their access rights (roles). According to the access rights (role), the data accessible to users will vary and to accommodate that, web UI have different views. The application is based on a boilerplate for Node.js web applications called "Hackathon Starter". This boilerplate uses a template engine called Pug which allows for rapid development of dynamic reusable content. The Pug template engine compiles the Pug code to HTML at compile time. The benefit of using a template engine is that it allows reusing static web page elements, while defining dynamic elements based on the data.

The data can be accessed through the Web UI in two data formats - raw data using direct SPARQL queries and data modified. The data analytics are presented by role through the use of dashboards. The web application has four separate views for each user and the data varies according to the user. Figure 2.5 shows an example of a dashboard for a cloud provider.

2.3.3 Data market encryption

The encrypted file system work utilizes two tools developed by the University of Bergamo, namely; the 'FreyaFS' encrypted file system which leverages the 'aesmix' all-or-nothing transform Mix&Slice encryption library. We carried out three primary modifications so far: the containerization of the 'FreyaFS' encrypted file system, the investigation of container state externalization and the in-

Figure 2.4: UC1 Android Auto interface after selecting policy for electricity cost per Kw

tegration of external key management. The containerization allowed for a integration with the existing MOSAICrOWN governance framework. Exploratory work to build upon the utilization of container volumes was carried out by leveraging an external object storage platform to facilitate container migration by making state-essential application data (such as 'FreyaFS' metadata files) remotely accessible to containers. Extending the 'FreyaFS' utility with an OASIS Key Management Interoperability Protocol client using PyKMIP allowed the replacement of the integrated password generator in 'FreyaFS' with a remote key management server which provides cryptographic keys and management functionality to 'FreyaFS', and more specifically the 'aesmix' cryptography library running in the background. 'FreyaFS' being written in the Python language meant the PyKMIP extension could be included directly into the code. The modified architecture of 'FreyaFS' is presented in Figure 2.6.

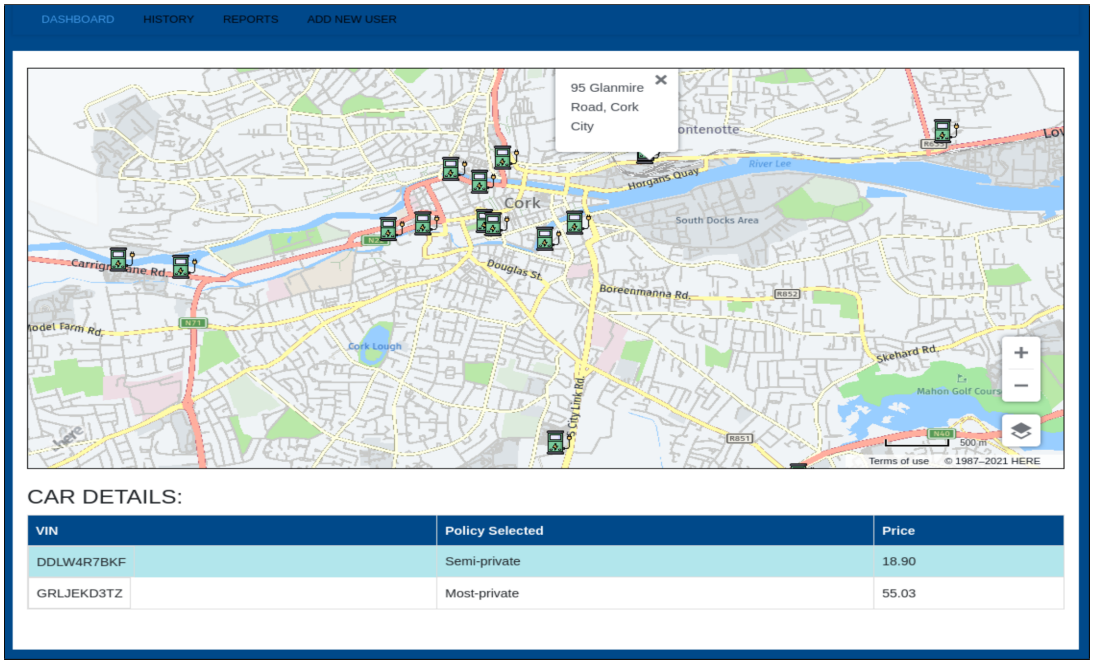


Figure 2.5: UC1 Dashboard for MOSAICrOWN Cloud Provider

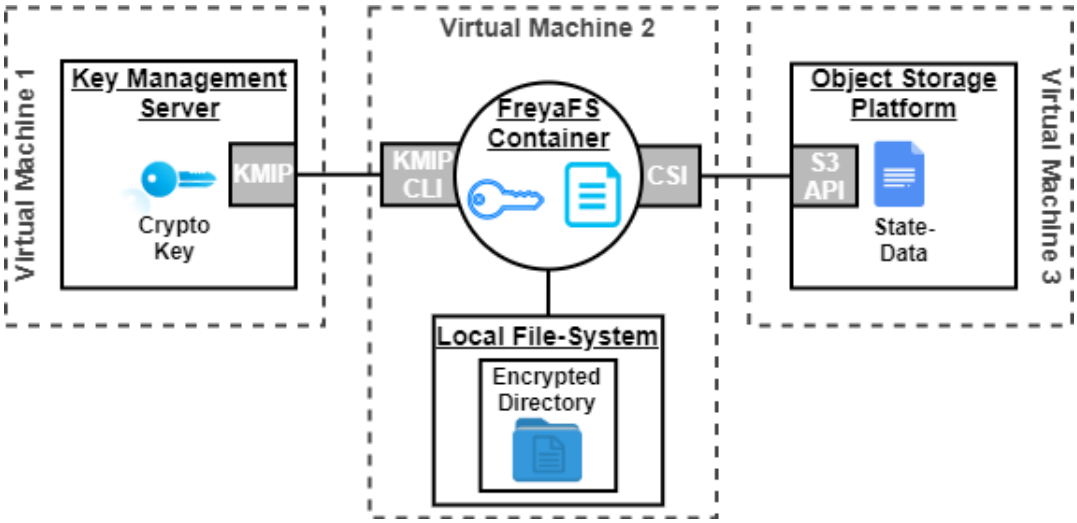


Figure 2.6: UC1 Architecture of modified FreyaFS utility

2.4 Analysis of the monitoring results

We present our analysis of the monitoring results in relation to the primary categories of requirements. These are data ingestion, data governance, access control management, data management, and data processing requirements in mind.

Data ingestion. Significant work has been carried out on the topic of data ingestion. The automotive tool presented in this deliverable addresses the need for some mechanism to ingest data from the EV. The governance framework delivered in D3.2 facilitates ingestion of data from the EV and also from the charging stations. The metadata model delivered in D3.1 gives context and vocabularies for the data fusion of policy, from D3.3, with various data formats.

Data governance. The work developed in T3.2 addresses some of the data governance re-

quirements. Combined with the customer web application developed for the deployment platform, and the ingestion of policy allowed for by the governance framework and the automotive application, then a large number of UC1 data governance requirements are satisfied.

Access control management. D2.2 describes potential solutions for access control management. Broadly speaking, these methods of enabling access control were adopted. Furthermore, the access control and roles introduced as part of the customer web application enable data sharing combined with the existing techniques from D3.2 facilitate the requirements surrounding access control.

Data management. As discussed in D2.2, features such as data at rest and data in transfer have been implemented, data at rest via the encrypted file-system presented in this deliverable. Combined with the features already existing within the deployment delivered to this point, a significant number of requirements have been satisfied at this point.

Data processing. A certain amount of functionality related to data processing has been added to the MOSAICrOWN deployment in the form of the analytics functions which are integrated into the customer web application. This processing populates the dashboards which are presented to the customers and can be in the form of aggregate data.

We list the functional requirements and the components addressing them in Table 2.1.

Requirement Reference	Description	Dimension	Covered by Component
REQ-UC1-DI1	MOSAICrOWN ingestion mechanism should support close to source deployment	Ingestion, Policies	Android Auto application
REQ-UC1-DI2	Ingestion mechanism should support real-time stream data handling	Ingestion	Android Auto application
REQ-UC1-DI7	Identifiers (e.g., VIN) should be preserved but secured from unauthorized access.	Storage, Policies	Encrypted file system
REQ-UC1-DI10	Ingestion mechanism should support ingestion from multiple concurrent sources.	Ingestion	Governance Framework, Android Auto application
REQ-UC1-AC6	The platform should allow data providers to share data with multiple data consumers.	Storage, Policies, Analytics	Governance Framework, Web UI
REQ-UC1-DM5	Data are protected at rest and in transfer.	Ingestion, Storage, Wrapping	Governance Framework, Encrypted file system
REQ-UC1-DE2	Distinction should be clear for consumers of the platform between those requesting data analytics functions and data sharing functions.	Storage, Policies, Analytics	Governance Framework, Web UI

Table 2.1: Use Case 1 requirements and their coverage by the components by tools provided in the first version of tools.

2.5 Findings

The research carried out to date in relation to UC1 has satisfied a significant number of the requirements from D2.1. The governance framework described in D3.2 has been enhanced by the automotive tool which was introduced in Section 2.3 and is described in detail in W2.3 (“Tools for CTI platforms”). Furthermore, the customer web application also aligns with work carried out in WP3 related to data governance and data management.

The tools implemented as part of WP4 have also facilitated UC1 with regard to data wrapping. Through the extension of these tools with containerization technology, container state externalization and the integration of external key management we have further satisfied the requirements UC1 has in order to deliver services and data protection presented in D2.1.

To complete T2.3 and deliver a complete use case prototype, which will be described in D2.4, we need to integrate either tools or techniques from WP5 if appropriate. Ideally, the tools developed in WP5 could be integrated, however we will investigate others if necessary. Also, the datamarket filter needs to be evaluated further to determine the best approach such that the filter behaves according to the policies stored within the data lake.

3. Use Case 2 (MC)

Use Case 2 (UC2) considers of financial institutions in the context of confidentially and confidently sharing data. While it may seem obvious that financial data includes varying levels of personally identifiable information (PII), the less obvious aspect of UC2 is determining how to define, store, wrap and, eventually, analyze this PII. Without the ability to securely store, access and utilize this data, financial institutions lose a key part of their revenue and innovation streams, with a critical impact in their credibility.

Data Governance, Wrapping and Sanitization are key components of UC2 - without any of these three steps, the data financial institutions interact with on a daily basis would be unusable and potentially pose a threat to their clients and cardholders.

As each of the work packages detail, UC2 poses interesting caveats to the stereotypical and usual approach to each component. Given the high visibility of transaction-level data and the amount of PII tied to every transaction, it truly is critical to ensure the data resulting from every swipe, tap, or click is securitized and treated with the importance it demands.

3.1 Technical overview

As the regulatory and compliance landscapes change and data is continuously shared amongst multiple parties, the usage of personal data has increasingly narrowed and shifted power to the customer.

Businesses now need to obtain explicit consent for specific usage and access of personal data. In the advent of GDPR (and other similar privacy regulations in the world) and these new restrictions, governments, commercial enterprises, and charitable organizations have been reluctant to share data.

While Mastercard, business owner of UC2, has a business need to operate on combined data provided by multiple parties to analyze at both microeconomic and macroeconomic levels, all financial institutions have an increasing need to use data to leverage their business. Furthermore, with the advent of open banking, there is a core need to pool data from all players in the ecosystem to enable new and improved product offerings, analysis on customer needs, and utilization of current products.

It is important to note that it is paramount to protect the privacy of individual data contributor's information, but it is also critical that the combined data asset be protected with the same rigor as well, since the access to combined information can be used to gain undue market level advantage. There is further need to track the origins and lineage of the data in the data market life-cycle. The development of novel techniques for providing effective data protection enables better and enriched data analytics.

Figure 3.1 describes the MOSAICrOWN dimensions that apply to Use Case 2. MOSAICrOWN can provide solutions for enabling the sharing and processing of microdata in respect of privacy regulations and on possible privacy/confidentiality constraints holding over the data.

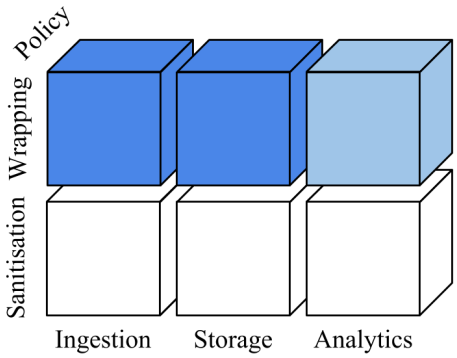


Figure 3.1: Overview of Use Case 2 dimensions

These scenarios require techniques, provided by data wrapping, which are first operated by the data owner (ingestion phase), and by the data analyst and data processor at the storage and analytics phase (in case of internal processing) or at the storage phase only (in case of data extraction for external analytics).

3.2 Monitoring of the tasks

Use Case 2 is concerned with WPs 3 through 5, covering a variety of components such as storage, security and sanitization of financial data so it is safe and ready for analysis.

3.2.1 WP3 monitoring and alignment

WP3 provides multiple examples of data governance framework and policy language to regulate the access and use of data for UC2 data. The policy language configuration file translates the MasterCard interpretation of the GDPR in wrapping techniques to be applied to specific data type. The platform can manage different privacy regulations (GDPR, LGPD, POPI,..); for each of them a configuration file in JSON format is created and uploaded into the system.

WP3 includes T3.1, T3.2 and T3.3 covering respectively reference metadata model, policy model and languages and finally policy management.

WP3 is strictly connected with UC2, providing the policy languages that is one of the inputs for the UC2. In fact, all wrapping techniques applied to data ingested and managed in UC2 is defined according with the privacy regulation applicable and defined by the policy languages.

The policy language template associate aith each type of data the suggested wrapping techniques, mainly impacted by privacy law and data type and distribution. Different privacy laws (i.e. LGPD and GDPR) could consider, from a privacy perspective, different levels of sensibility for the same field (i.e. address). This means that different wrapping techniques are suggested and applied.

3.2.2 WP4 monitoring and alignment

WP4, the main research effort of MC for MOSAICrOWN, outlines the various forms of data wrapping that are available for use. Different classifications of data could and should require different wrapping techniques. The platform is able to recognize the different data type thanks to its

integrated proprietary semantic data type detection model. Starting from a dataset, the model automatically identifies the semantic types within each column of the input file leveraging a proprietary algorithm.

WP4 includes T4.1, T4.2, T4.3 and T4.4 covering respectively basic techniques of data protection, enforcement mechanism, controlled query execution and incentives and economic aspects. UC2 is the included in WP4 and provides an example of how wrapping techniques are applied to a financial data set, according with the selected privacy regulation (modelled into a privacy language template).

UC2 is extensible to any type of industry and market: data is collected in unstructured way and analyzed and interpreted by a cloud machine learning model, able to understand data type, semantic meaning, and data distribution. All those attributes define the metadata type and is an input for the application of the right wrapping technique.

3.2.3 WP5 monitoring and alignment

WP5 focuses mostly on data sanitization integrating the wrapping techniques investigated in WP4. The key component of WP5 that is relevant to UC2 is the ability to guarantee security regardless of access level. The model described enforces a multi-tiered access structure that best fits UC2. This means that, depending on their role, users are able to access varying amounts of securitized data for analysis, successfully allowing financial institutions to continue to utilize transaction level data without putting the data at risk of breach.

WP5 includes tasks T5.1, T5.2 and T5.3 covering respectively privacy metrics and risks, data sanitization techniques and collaborative computation for sanitization. Furthermore, the work completed in WP5 allows for this model to be utilized in parallel with the data wrapping techniques described in WP4. This makes wrapping and sanitization simultaneous in the MOSAICrOWN workflow, an added benefit for UC2.

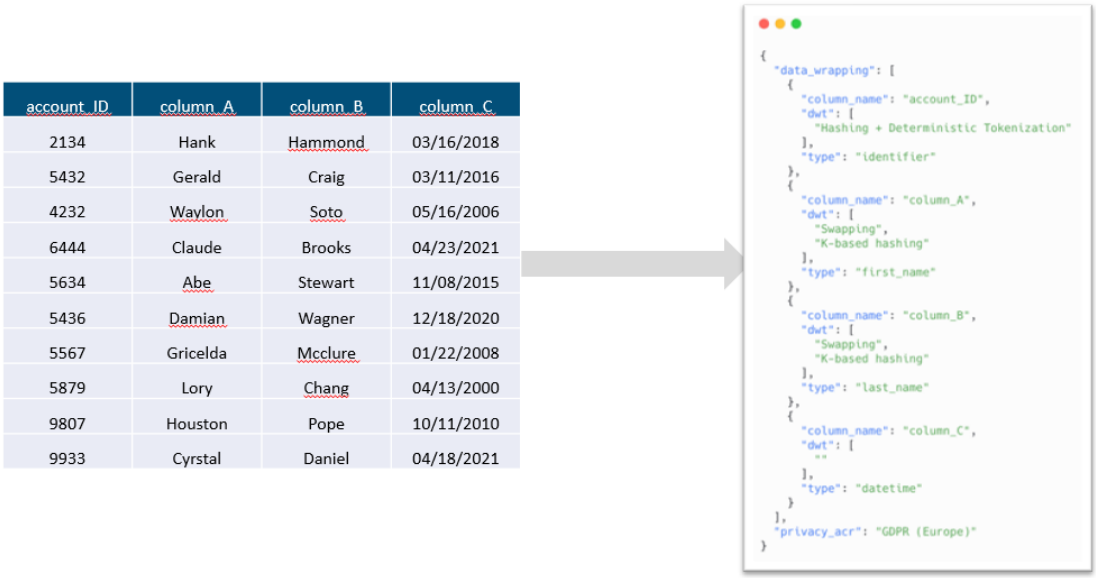
WP5 provides to UC2 the sanitization policies to apply, avoiding risk of personal data identification. In each step of the UC2 data, once used, PII are immediately destroyed. We consider two levels of wrapping, a first level happens once the data is wrapped by the customer and sent to the data market. The second level happens once the data market wraps the file provided by the organization. This way there is no connection between source files and the final wrapped data.

Additionally all temporary and working data is deleted step by step, and also on cloud data storage info are fragmented and duplicated, using a specific key.

3.3 Deployment status

Use Case 2 is a cloud-based platform to upload data files for anonymization. The user selects a policy language in alignment to the type of data and/or regulations to be applied. Then, the platform runs a semantic analysis of the data provided, recognizing data type, semantic and data distribution of the data set uploaded by the user. Accordingly, wrapping techniques to be applied are proposed to the user.

The wrapping techniques can be customized by the user selecting options into the list defined for each metadata by the policy language. Once the file is anonymized, the user can download the fully anonymized data set.



account_ID	column_A	column_B	column_C
2134	Hank	Hammond	03/16/2018
5432	Gerald	Craig	03/11/2016
4232	Waylon	Soto	05/16/2006
6444	Claude	Brooks	04/23/2021
5634	Abe	Stewart	11/08/2015
5436	Damian	Wagner	12/18/2020
5567	Gricelda	Mcclure	01/22/2008
5879	Lory	Chang	04/13/2000
9807	Houston	Pope	10/11/2010
9933	Cyrstal	Daniel	04/18/2021


```
{
  "data_wrapping": [
    {
      "column_name": "account_ID",
      "dwt": [
        "Hashing + Deterministic Tokenization"
      ],
      "type": "identifier"
    },
    {
      "column_name": "column_A",
      "dwt": [
        "Swapping",
        "K-based hashing"
      ],
      "type": "first_name"
    },
    {
      "column_name": "column_B",
      "dwt": [
        "Swapping",
        "K-based hashing"
      ],
      "type": "last_name"
    },
    {
      "column_name": "column_C",
      "dwt": [
        ""
      ],
      "type": "datetime"
    }
  ],
  "privacy_acr": "GDPR (Europe)"
}
```


Figure 3.2: UC2 Configuration file

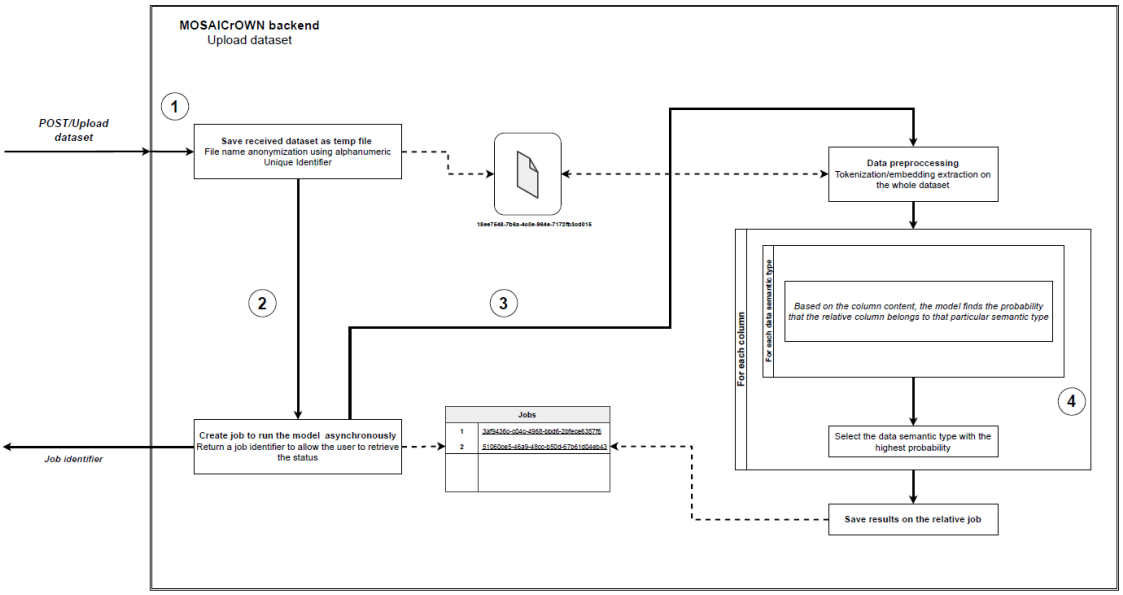


Figure 3.3: UC2 Process diagram

Figure 3.2 is an example of sample data set with the internal representation of the policy language file that illustrates the data type recognition process.

Figure 3.3 describes the overall process of data analysis and wrapping, starting from an unstructured data set provided by the user.

The process is fully asynchronous (mandatory considering the amount of data involved in this data transformation), so the user can run it and then retrieve the related job monitoring the execution or moving to the next step.

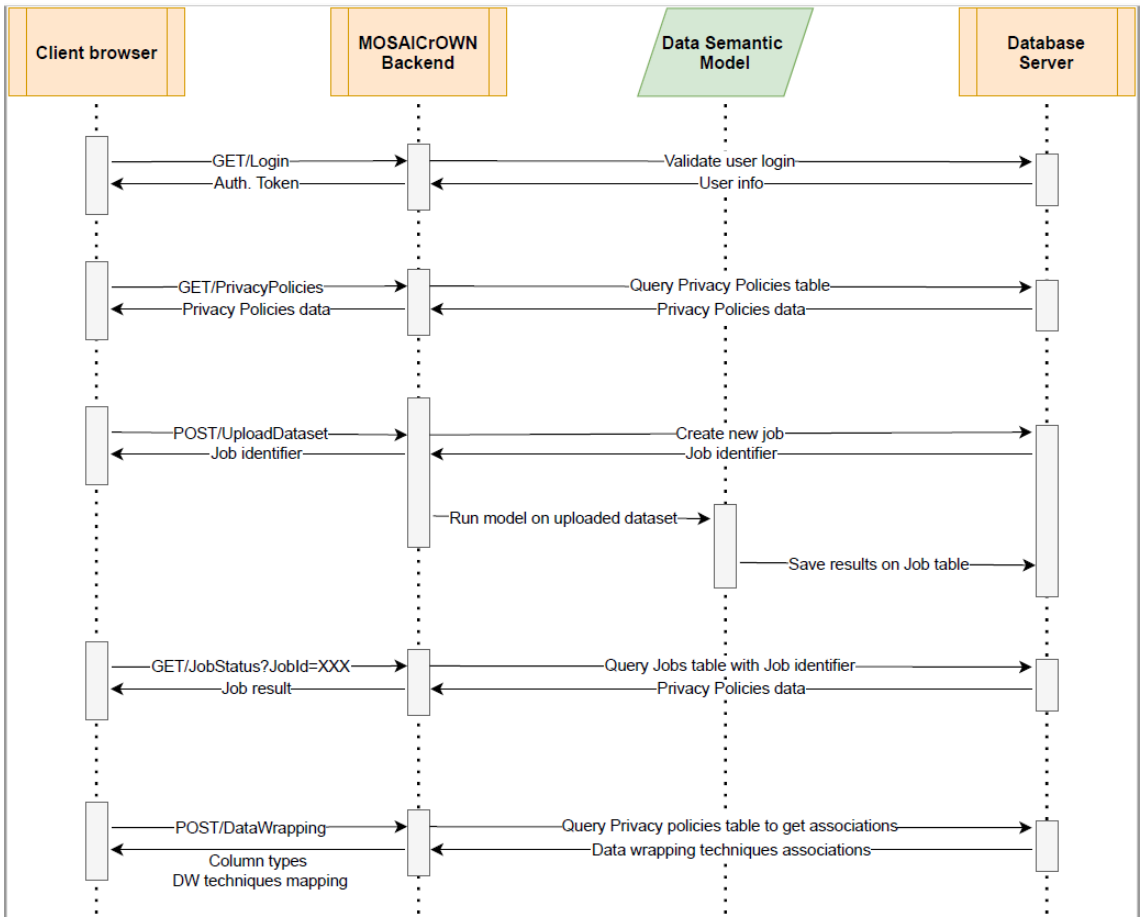


Figure 3.4: UC2 Sequence process

Figure 3.4 puts in evidence the different sequences we defined in the pilot, and the different components involved in the several layers of the platform.

3.4 Analysis of the monitoring results

3.4.1 Cloud storage

In UC2, cloud storage is considered as repository for all data, considering the high volumes of data involved in financial institutions analytics and insights generation. It supports both original data provided by the organizations and final wrapped information, including all temporary data generated and sanitized during the process.

Cloud storage allows extensibility to any type of data and volumes, and consequently to every industry and type of customers.

Data is secured implementing data encryption, fragmentation, and replication. Data is encoded using a private key, fragmented over the cloud and replicated to have maximum security in case of fault.

Any public cloud provider can be considered for UC2, so there are no specific indications about the vendor to use.

Requirements Reference	Description	Covered by component
REQ-UC2-DI1	Batch handling	Ingestion
REQ-UC2-DI2	Data types and formats	Ingestion
REQ-UC2-DI3	Data feeds via APIs	Ingestion
REQ-UC2-DI4	Data perturbation	Ingestion
REQ-UC2-AC4	API Access only	Access Control
REQ-UC2-W1	Data Risk Assessment	Data Wrapping
REQ-UC2-W2	Data Wrapping Approach	Data Wrapping
REQ-UC2-W3	Data Catalogue	Data Wrapping
REQ-UC2-W4	Wrapping Techniques	Data Wrapping
REQ-UC2-W5	Pseudonymizantion Data	Data Wrapping
REQ-UC2-W6	Re-anonymizantion for UIDs	Data Wrapping

Table 3.1: Use Case 2 requirements and their coverage by the components by tools provided in the first version of tools

3.4.2 Security

UC2 applies all levels of security required, from data storage to data transfer.

Indeed, regarding data storage security, files and encryption keys are encrypted on the client side before being stored in the cloud.

Additionally, the encrypted data transfer is done via a secured channel. The cloud provider never has access to the encryption key and consequently the provider cannot access the content of any stored file.

3.4.3 Sharing

Sharing outside of the organization of the protected data is a functionality that can be raised if required. It can be created in an area on the cloud storage where users can exchange encrypted files with other users of this service, defining an area for each user for this scope.

Also a public area can be considered to share data with all users in a public way.

3.4.4 Access

In UC2, the access is done by means of user authentication by using the email and password as credentials.

3.4.5 Analysis of Functional Requirements

The functional requirements for Use Case 2 prefixed with REQ-UC2 and are divided into three categories:

Ingestion (DI) requirements focusing on data ingestion.

Access Control (AC) requirements focusing on access management.

Data Wrapping (W) requirements focusing on data wrapping techniques.

We list the functional requirements of the current UC2 POC and the components addressing them in Table 3.1

3.5 Findings

UC2 provides a complete end-to-end approach to data preparation for analytics in a fully privacy compliant way. The process leverages on cloud capabilities not only for storage but also for the semantic analysis performed on the unstructurerd input data. According to regulation selected, the UC2 identifies for each metadata field all the wrapping techniques that can be applied, related to the level of privacy sensibility defined by the regulation. UC2 delivers two different outputs: one is the wrapped dataset, generated by a source provided and loaded into platform by the user, while the second is the policy language template, identifying the wrapping techniques applicable to each fields.

The wrapped data set is the source for all analytical products of the organization and generates insights supporting business and strategical decisions in a full privacy compliant way.

4. Use Case 3 (SAP SE)

The objective of Use Case 3 (UC3) is to enable privacy-preserving consumer analytics via a cloud-based data market. UC3 considers business-to-business (B2B) data sharing, potentially containing personal (customer) information as well as sensitive business data. There are valuable, holistic insights that can be gained from combining the distributed data hidden in data vaults of different companies. However, to ensure that the data can be shared adequate protections are required.

The technological challenges of this use case are concerned precisely with these protections. Different sanitization techniques are investigated and implemented in the course of MOSAICrOWN, as well as privacy metrics to quantify and bound the risks for such data sharing to an acceptable level while still providing a meaningful utility.

In this chapter, we will first provide a technical overview of Use Case 3 in Section 4.1, where we also describe the tools by which we achieve the requirements collected in Deliverable D2.1. In Section 4.2, we describe our activities on monitoring of relevant tasks from the work packages that contribute towards UC3. Subsequently, in Section 4.3, we put our focus on the deployment of the sanitization tool "DPtool", which supports the main functional requirements. In particular, we detail its architecture, and provide an overview of the application from the users' perspective. Finally, we analyze the results of the monitoring activities and the alignment of the functional requirements (Section 4.4.1).

4.1 Technical overview

Use Case 3 deals with sanitization of sensitive data during the data life-cycle – ingestion, storage, and analytics – as highlighted in Figure 4.1. It is a business-to-business (B2B) scenario, where different business partners (e.g., a producer of goods and retailer(s)) want to collaboratively perform privacy-preserving analytics over their combined data.

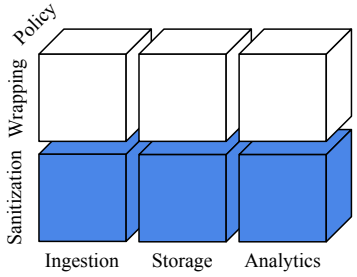


Figure 4.1: Overview of Use Case 3 dimensions

The data considered in this use case is of two types. It can be in the form of (sensitive) O-data, i.e., *operational* data; usually, key performance indicators such as sales numbers, cash-to-cash cycle time. Also, it can be (personal) X-data, i.e., *experience* data such as consumer

ratings, purchase histories. The goal is to redirect budgets to increase customer satisfaction (e.g., by allocating more money for certain marketing expenditure or joint promotional campaigns), and improve operational efficiency. In general, anything involved with the operational process: how efficiently it is run as well as the experience of potential inefficiencies (e.g., long delivery) by customers.

To protect these types of data and to realize Use Case 3 we provide tools for:

- sanitization techniques with strong formal privacy guarantees via differential privacy,
- interpretation of these guarantees to provide guidance for the parametrization of the sanitization to achieve desired protection and utility goals.

Differential privacy, a semantic privacy notion, restricts what a (randomized) mechanism operation on sensitive data can output. The randomization is mainly achieved via perturbation in form of additive noise. When the input data changes in a single record, the change in the output (distribution) is bounded by the privacy parameter ε. Data sets that differ in a single record are called *neighbors*. Informally, a differentially private mechanism M provides a form of indistinguishability for neighboring data sets D and D', as one cannot distinguish with certainty which one was the input of M based on the output of M.

Differential privacy supports different models of implementation as described next.

Local model: the data is sanitized at ingestion and analytical evaluations are performed over the already sanitized data.

Central model: the analytical evaluations are performed on unsanitized data and the result of the evaluations are sanitized instead.

Hybrid model: combines the local and central model by sanitizing the result of a computation (as the central model) without requiring plaintext access (like the local model) via cryptography.

4.1.1 Tools to realize UC3

To achieve the goals of UC3, we provide the following tools for data sanitization, as well as a tool for the interpretation of privacy guarantees and guidance on choosing privacy parameters:

DPtool provides multiple interfaces exposing differential privacy mechanisms which can be applied on data sets to sanitize them during ingestion (i.e., only sanitized data is stored) or during analytics (i.e., data is stored in plaintext or encrypted). On one hand, DPtool provides an intuitive graphical interface realized with SAPUI5. On the other hand, DPtool also offers a programmatic interface in the form of a REST API allowing automatic sanitization. The SAPUI5 GUI provides manual sanitization and supports data analyst by evaluating different mechanisms. We will extensively describe DPtool in the remainder of this chapter.

MIAtool supports data scientist in the parameterization of differential privacy in the context of machine learning. MIAtool can be applied during ingestion or investigation of selected analytical functions. MIAtool has a graphical user interface also based on SAPUI5. MIAtool assists analysts and data scientists to evaluate the guarantees of differential privacy. For this goal, membership inference risks are computed and it provides a measure of the utility-privacy trade-offs due to the parameterization. We provide a detailed description of MIAtool in D5.1 (First version of data sanitisation tools).

DPsc provides secure computation over two distributed data sets (as envisioned in the use case), i.e., allowing protected statistical inference without having to share any data at all. DPsc supports differentially private rank-based statistics (i.e., min, max, median). With this hybrid approach, combining differential privacy and secure computation, the ingestion and storage phases are skipped and the analytical phase is computed securely. DPsc is a command-line tool requiring mainly the input data and privacy parameter as parameters. A prototype implementation of DPsc will be described in D5.4 (Final versions of tools for data sanitisation and WP5 computation).

While monitoring of the tasks will concern the overall aspects of Use Case 3, in the following, we focus on the description of the sanitization tool "DPtool", as it supports the main functional requirements of the use case. Particularly, we will detail its architecture (Section 4.3.1), and provide an application overview (Section 4.3.2).

4.2 Monitoring of the tasks

Regarding the identified requirements detailed in D2.1, Use Case 3 (UC3) is concerned with WP3, WP4, and in particular WP5. Therefore, our monitoring activities are focused on those tasks of WPs3–5.

4.2.1 WP3 monitoring and alignment

The data governance framework, defined and developed in the course of WP3, aims to support sophisticated data protection with a description language usable by non-experts. Ease of use for non-experts w.r.t. data protection guarantees and techniques is an important goal in UC3 to attract and support business customers, e.g., the retailer and producer mentioned in UC3.

For this, SAP SE actively participated in discussions and descriptions found in a first version of the reference metadata model (T3.1), and will closely follow the ongoing efforts in the creation of the policy model and language (T3.2) and policy management (T3.3). The results of WP3, namely policy specifications for data protection mechanisms, are also aligned with the mechanisms defined in WP4, which we describe next.

4.2.2 WP4 monitoring and alignment

Data wrapping is a security mechanism that can be used additionally to sanitization (a non-reversible randomized data transformation aiming to preserve statistical insights), which is the main goal of UC3. Security mechanisms identified during WP4 include, e.g., hashing (non-invertible and deterministic, i.e., the same input is always mapped to the same output) and encryption (non-reversible without additional knowledge, i.e., a decryption key). Basic techniques (T4.1) and enforcement mechanisms (T4.2) for such additional data protection can be useful for UC3 actors, as they provide an additional protection layer. Furthermore, WP4 techniques and mechanisms might enable the actors to ingest the data in an encrypted fashion (instead of plaintext as described for the central model) and only allow (partial) decryption during an analytical processing phase. Collaborative analytics in the hybrid model rely on cryptographic tools similar to encryption (namely, secure computation) to allow UC3 actors to jointly compute an analytical function over their data without sharing it with either the cloud service provider or other actors.

4.2.3 WP5 monitoring and alignment

WP5 investigates data sanitization, the main research effort of SAP SE in the course of MOSAICrOWN. SAP SE is leading this work package and it provides the main components to realize UC3. WP5 covers privacy metrics and risks (T5.1), which enable UC3's business customers to assess and bound potential privacy risks, to make informed decisions and provide guidance for the parameterization of privacy-utility trade-offs. Another major aspect for UC3 are data sanitization techniques (T5.2) for different data types (e.g., simple numbers and unstructured text) that support strong, formal privacy guarantees and yet aim to provide high utility. Additionally, collaborative computation for sanitisation (T5.3) covers techniques to realize the hybrid model via secure computation of sanitization mechanisms. This enables participation in a data market even for UC3 actors which do not wish to persistently store their data or reveal any parts of it to the cloud service provider even in the course of an analytical function evaluation.

4.3 Deployment status

This section consists of two main parts and provides a detailed look at the key components of DPtool. In the first part, we provide an overview of DPtool's architecture and the way how components interact with one another. In the second part, we provide a user-centered point of view of DPtool and describe its user interface and usage, as well as the supported anonymization methods.

4.3.1 Architecture

First, technical details of DPtool, which include implementation decisions such as utilized technologies and frameworks, are described. Second, the main component performing anonymization tasks is covered. Next, the backend system, which handles, e.g., the storage of data, is described. In the end, a complete overview of DPtool's architecture is given and the way how components interact with one another is explained.

SAPUI5 GUI

The graphical user interface of DPtool is implemented based on SAPUI5 (SAP User Interface for HTML 5) a framework providing functionality to develop mobile applications. In particular, SAPUI5 enables the possibility to build web applications according to HTML5 development standards [SAP0Za]. SAPUI5 by design provides multiple javascript libraries making development more efficient. Additionally, it encourages the developer to use the Model-View-Controller (MVC) concept by structuring project files accordingly. Model and controller are based on javascript while views are designed for example with HTML or XML.

For DPtool the views are designed by means of XML since it is a human-readable markup language which does not need any additional software libraries and is supported by all modern browsers. Thus, maintainability and usability are ensured. The controller encapsulates functionality enabling the user to ingest data and configure multiple data anonymization methods. For internationalization (aka localization) and to simplify configuration, texts are stored in i18n property files separate from the UI. Hence, translating texts into different languages is easier [SAP0Zb]. Thus, it is possible to extend DPtool to support multiple languages. By now, the prototype only

supports English. Last, SAPUI5 web applications by default support responsiveness making DPtool usable on different devices and screen formats.

REST-API

The REST-API is at the core of DPtool. REST stands for REpresentational State Transfer and provides an easy to use interface to fetch data and execute commands. Specifically, the REST-API of DPtool provides multiple anonymization methods and makes them accessible for programmatic access, i.e., in the form of scripts.

The current API is based on earlier research efforts that were done as part of the EU H2020 C3ISP project (cf. C3ISP D8.3, Section 8.1). The API in C3ISP was designed for specific applications (threat intelligence sharing), with limited parameterization (e.g., sensitivity fixed to 1 to anonymize counts with Laplace mechanism). We updated the API to support more use cases, improved parameterization, and extended it with additional mechanisms (Gauss) as well as convenience functions for analytics (DP average, DP sum). Furthermore, a graphical user interface, as described earlier, was added to improve the usage experience for citizen developers, data scientists and analysts. The current version allows more control over the privacy parameters (cf. the description of the anonymization methods in Section 4.3.2) and hence is applicable to a wider and less specific variety of applications.

From a development point of view, the API is implemented with Spring framework. Spring is a Java based framework which is organized in about 20 modules intended to ease the development of web applications and predominately supports the implementation of restful services [Spr0Z]. Anonymization methods themselves are also implemented with Java making the integration into the REST API more efficient and smooth. Additionally, Spring is modular, allowing to import only task specific modules. Hence, applications can be extended more easily [Pac0Z]. The REST-API of DPtool is extended by Swagger which is an open source tool to simplify API documentation and development [Swa0Z].

PHP backend

Ingesting data via SAPUI5 GUI is handled by a PHP backend. PHP is a server side scripting language. For DPtool the PHP backend waits for an incoming request submitted via the SAPUI5 application. This request contains meta-data handling the selection and parameterization of anonymization methods for a given dataset. Both request and dataset are forwarded to the REST-API which returns the anonymized data. PHP logs several details of the anonymization request, for example applied methods or parameters and provides the log as well as the anonymized data as a report which can be downloaded.

Architectural view

Figure 4.2 shows the architecture of DPtool with its three components. The general workflow is as follows. First, data which is intended to be anonymized is ingested to the PHP backend. Then, an anonymization request is triggered and forwarded to the REST-API by PHP. Finally, the REST-API performs the requested anonymization, and the resulting data as well as the report are returned to the SAPUI5 frontend.

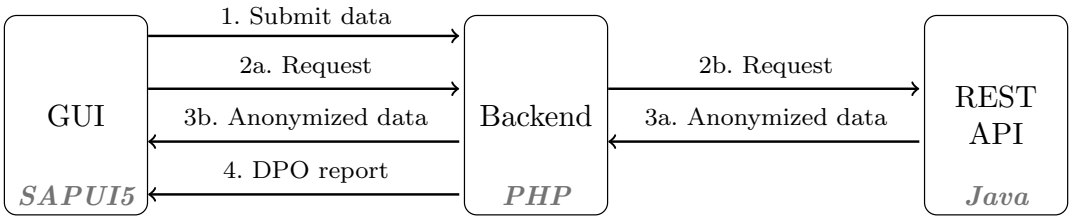


Figure 4.2: Architecture of DPtool

4.3.2 Application Overview

This section describes DPtool from a user-centered point of view. First, the user interface and its properties are introduced. Afterwards, the process on how to use the prototype is elaborated. Last, all supported anonymization methods are listed and discussed.

User interface and Usage

In order to ensure a good user experience, the user interface is kept simple to make the usage efficient and intuitive. A so-called file uploader implements the functionality to submit data. The uploader itself consists of two parts: 1) a button which opens a dialogue to browse for and select a CSV file when pressed, and 2) a text field which displays the absolute path to the selected file. After ingesting a new file, a radio button and a label with the name of the submitted file are automatically added beneath the file uploader. In case multiple files are ingested, selecting one specific file for further processing is performed by clicking the corresponding radio button. A table placed in the center of the GUI dynamically displays the features of the data which was ingested before. Features in this context denote the headers in a file. For every feature, one row is added to the table. If multiple files were submitted, the table refreshes accordingly to the selected radio button. Each feature (i.e., table row of the GUI) includes a name and a drop down menu which displays a list of anonymization methods. By hovering over one specific method, a tooltip pops up with a brief description of the particular method. If one method is selected, additional textfields or drop down menus are placed in the table row depending on the method implementation. These additional input fields are used to parameterize the methods. After choosing anonymization methods and parameters, the anonymization process can be triggered via a button located beneath the table. If no errors occurred during anonymization, a dialog appears showing the first five entries of the original file and the first five entries of the anonymized file as a preview. By pressing either a download button or a cancel button, the anonymized file can be fetched or discarded, respectively. Last, a DPO (Data Protection Officer) report can be downloaded by clicking a button placed at the bottom of the GUI. The report contains information on the selected anonymization method per feature, parameters, hash values of the original and anonymized columns, hash value of the entire anonymized dataset, applied hashing algorithm and a timestamp.

To start the workflow the user of DPtool ingests his or her data file(s) which should be anonymized. Afterwards, a specific file has to be selected by clicking the particular radio button. For each feature at most one anonymization method can be selected. If needed by method design, additional input fields are displayed which have to be set to perform anonymization. As soon as the user has chosen all methods and set all parameters, the user triggers the anonymization

task. A preview of the original and anonymized dataset is displayed as described above and the anonymized file can be fetched. Last, the user gets a DPO report summarizing the key elements of the entire anonymization process.

Anonymization methods

This section covers the methods of DPtool and gives a description of each method. We provide methods for *suppression* (via removal), *perturbation*, i.e., additive noise as used in the Laplace and Gaussian mechanism, and *probabilistic selection* as used in the Exponential mechanism for privately computing the median. For convenience, we also provide methods for differentially private summation and averaging based on the Laplace mechanism. Additionally, location data can be anonymized via *geo-indistinguishability* (GI) [ABCP13]. Informally, GI, a generalization of differential privacy, ensures that it is more likely to output coordinates close to a location than far away points. More formally, for a GI mechanism M, the distance of the output distribution (denoted d_{distr}) between two locations x, y is bounded by their Euclidian distance $d(x,y) \leq r$ such that $d_{distr}(M(x), M(y)) \leq \varepsilon \cdot r$.

Methods that fulfill differential privacy in general rely on the privacy parameter `epsilon`, where some might additionally require a second parameter `delta` and the `sensitivity` to be specified. For usability, some methods provide a pre-defined selection of anonymization parameter choices which we call `ProtectionLevel`.

Note that the current API names and functionality are subject to change due to ongoing development. The following anonymization methods are provided by DPtool and require all a CSV `columnName` as parameter (to determine where to apply the anonymization).

Removal: The entire attribute is removed (anonymization by suppression), e.g., for personal identifiers such as ID numbers or full name.

Removal up to delimiter: Removes substrings (beginning from the left or right) up to a delimiter (e.g. '.' for IPv4 addresses, ':' for MAC/IPv6, '@' for email).

- `Area`: LOWER/UPPER removal begins from the left/right.
- `Delimiter`: Character, e.g., IPv4 addresses (.), for MAC/IPv6 (:), email (@).
- `ProtectionLevel`: LOW/MEDIUM/HIGH/FULL removes up to the 1st/2nd/3rd/4th delimiter.

Laplace Mechanism: Adds Laplace distributed noise to query results with given sensitivity (e.g. 1 for count queries) and privacy parameter ε.

- `Epsilon`: Float (privacy parameter of differential privacy).
- `Sensitivity`: Float (maximum impact that the inclusion/exclusion of an individual can have on a query / function evaluation).

Gaussian Mechanism: Adds Gaussian distributed noise to query results with given sensitivity (e.g. 1 for count queries) and privacy parameters ε and δ.

- `Epsilon`: Float (privacy parameter of differential privacy).
- `Delta`: Float (privacy parameter of differential privacy, ideally this should be negligible in the number of records).

- `Sensitivity`: Float (maximum impact that the inclusion/exclusion of an individual can have on a query / function evaluation).

Laplace Sum: Convenience function which calculates the sum of values that have been perturbed with Laplace noise (see above). The sensitivity (automatically computed over the data, assumed to contain min and max from domain) is the maximum impact that the inclusion/exclusion of an individual (record) can have on the result of the summation.

- `Epsilon`: Float (privacy parameter of differential privacy).

Laplace Average: Convenience function which calculates the mean over values that have been perturbed with Laplace noise (see above) within bounded ranges.

- `Epsilon`: Float (privacy parameter of differential privacy).
- `a_min`: Float (lower bound of data range).
- `a_max`: Float (upper bound on data range).

Latitude: Provides Geo-Indistinguishability for locations, we provide one method for latitude and one for longitude as these values are commonly separated in CSV.

- `Epsilon`: Float (privacy parameter of differential privacy).

Longitude: See **Latitude**.

Exponential Mechanism for Median: The Exponential Mechanism enables Differential Privacy without perturbation for functions (e.g., median), by selecting the element with highest utility, i.e., closest to the actual result, from the function output range with high probability (informally, a selection probability is defined for all elements, good results are exponentially more likely).

- `Epsilon`: Float (privacy parameter of differential privacy).

Since DPtool deals with sensitive data intended to be anonymized, DPtool itself needs to ensure that data is kept safe. Therefore, DPtool provides encryption during data transmission in order to make network traffic secure via HTTPS. Additional secure storage, in the form of encryption-at-rest can be realized via tools from WP4.

4.4 Analysis of the monitoring results

The tools for Use Case 3 are built to fulfill the requirements defined in D2.1 “Requirements from the Use Cases”. We provide an analysis of the functional requirements in Section 4.4.1. The sanitization for Use Case 3 fulfills Differential Privacy as detailed in Deliverable D5.2 “First report on privacy metrics and data sanitisation”. Privacy quantification, in the context of machine learning, is provided via Membership Inference Attacks (MIA); the theoretical details of Membership Inference are also detailed in D5.2 “First report on privacy metrics and data sanitisation” and a first tool for Membership Inference, MIAtool, is described in Deliverable D5.1 “First version of data sanitisation tools”. Details on the DPsc tool based on secure multiparty computation (MPC) will be provided in the upcoming Deliverable D5.4. A first version of the meta-data, describing what information and parameters are required to apply the sanitization, is provided in Deliverable D3.1 “First version of the reference metadata model”.

4.4.1 Analysis of Functional Requirements

The functional requirements for Use Case 3 – prefixed with "REQ-UC3-" – are divided into three categories:

Access control ("AC") management, mainly provided by the tools from WP3 (also existing solutions, e.g., from database systems such as SAP HANA could be used).

Local sanitization ("SL") deals mostly with parameterization and execution of the sanitization; additionally, it covers the storing and sharing of the anonymized result.

Central sanitization ("SC") is primarily concerned with parameterization and execution of the sanitization; cryptographic techniques are also employed to support the hybrid model.

We list the functional requirements and the components addressing them in Table 4.1. Access control requirements (REQ-UC3-AC1, REQ-UC3-AC2, REQ-UC3-AC3) are not listed as they are met by existing solutions (e.g., SAP HANA).

Requirement Reference	Description	Dimension	Covered by Component
REQ-UC3-SL1	Local anonymization parameters chosen by the data owner	Ingestion, Sanitization, Policies	DPtool
REQ-UC3-SL2	Storing the result of the sanitization service in the Cloud	Ingestion, Sanitization	DPtool
REQ-UC3-SL3	Sharing sanitized results with other data owners	Storage, Policies	Data market
REQ-UC3-SC1	Central anonymization parameters chosen by the data owner	Analytics, Sanitization, Policies	DPtool, MIAtool
REQ-UC3-SC2	Collecting inputs from multiple data owners for aggregation	Analytics, Sanitization	Data market, MPC research prototype in upcoming D5.4 (DPsc)
REQ-UC3-SC3	Collecting inputs from multiple data owners via secure computation	Ingestion, Analytics, Sanitization	MPC research prototype in upcoming D5.4 (DPsc)
REQ-UC3-SC4	Anonymization to protect identity of data subjects and hinder re-identification	Ingestion, Analytics, Sanitization	DPtool

Table 4.1: Use Case 3 requirements and their coverage by the components by tools provided in the first version of tools.

Our sanitization tool DPtool already covers most sanitization requirements, however, the cryptographic tool supporting the hybrid model (REQ-UC3-SC3) requires additiona integration and implementationl efforts within the context of UC3. Currently, the functionality is provided by a research prototype, which we aim to make more user-friendly and further optimize for integration and application within Use Case 3. This is due to the complexity of these cryptographic tools, which are under active research and development (listed in Deliverable D2.2 "Report on requirements, research alignment and deployment plan"), and their additional overhead (in the form of

computational complexity as well as software dependencies). As mentioned above, the hybrid prototype will be described in detail in the upcoming D5.4.

4.5 Findings

Our tools fully cover the requirements to satisfy Use Case 3, i.e., enabling business customers to perform privacy-preserving analytics, as we have confirmed in Section 4.4.

Our sanitization tool DPtool provides a programmatic REST API as well as an intuitive graphical user interface, accustomed to users of SAP SE's business applications, and supports the main differential privacy mechanisms.

To complement DPtool, we have developed MIAtool, which provides the Membership Inference Service as described in Deliverable D5.1 "First version of data sanitisation tools". With MIAtool, we cover the interpretation and guidance w.r.t. the privacy guarantees (non-functional requirement REQ-UC3-IN1, i.e., Anonymization services should be accompanied by simulated adversary).

Research prototype DPsc, which we will detail in Deliverable D5.4, implements the hybrid model based on secure computation techniques (garbled circuits [Yao86, BHR12] and secret sharing [Bla79, Sha79]).

5. Conclusions

This document is the final report on research alignment of the use cases development, which intermediary statuses were shared in D2.1 and D2.2, this in line with the WPs 3 through 5 and provides the final report on research alignment on the use cases at the month 30 of the project. Each chapter covered a different use case of MOSAICrOWN identified by their assigned industry partners. The use cases span from ICV data protection in intelligent connected vehicle (UC1) to financial institutions and their transaction-level data anonymization (UC2) to a cloud-based consumer-centric data market (UC3).

MOSAICrOWN supports multiple deployment options by considering where the protection techniques should be applied: data sanitization and wrapping locally, a central platform providing these functionalities as a service during ingestion and the entire data life-cycle, as well as hybrid scenarios, combining both previously mentioned scenarios, which can be augmented by cryptographic tools, e.g., secure computation. Then, to meet security and privacy requirements for the use cases, MOSAICrOWN provides different wrapping as well as data sanitization techniques to augment policy-based protection mechanisms which control data access, usage and sharing. Finally, to provide meaningful utility while providing strong security and privacy guarantees special care has to be taken with regards to how to sanitize the data to protect the privacy of individuals while allowing statistical inference. For wrapping techniques enforcement of policies and appropriate wrapping technique according to data type recognition are the key for efficient data protection and yet enable to perform analytics over the protected data matching regulation requirements.

Early versions of the tools are already available, the final collection will be prepared before the end of the year. D2.4 “Use case prototype” covering task T2.4 on the final testing and use cases validation will be due at month M36. The monitoring is also helping to build the planned use of the results per use case in terms of applicability, exploitation and also on joint exploitation activities that will be carried out in WP6.

Bibliography

[ABCP13] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications security*, pages 901–914, 2013.

[BHR12] Mihir Bellare, Viet Tung Hoang, and Phillip Rogaway. Foundations of garbled circuits. In *Proceedings of the annual ACM conference on Computer and Communications Security*, 2012.

[Bla79] George Robert Blakley. Safeguarding cryptographic keys. In *International Workshop on Managing Requirements Knowledge*. IEEE, 1979.

[Pac0Z] Packt. Benefits of the spring framework - learning spring application development, 2020-09-03T10:41:06.000Z.

[SAP0Za] SAP. What is sapui5? - definition from whatis.com, 2020-09-03T10:40:07.000Z.

[SAP0Zb] SAP. Sapui5 translation tables, 2020-09-03T10:40:27.000Z.

[Sha79] Adi Shamir. How to share a secret. *Communications of the ACM*, 1979.

[Spr0Z] Spring. Introduction to spring framework, 2020-09-03T10:40:46.000Z.

[Swa0Z] Swagger. Swagger api documentation & design tools for teams, 2020-09-03T10:41:17.000Z.

[Yao86] A.C.-C. Yao. How to generate and exchange secrets. In *Proc. of IEEE FOCS*, Toronto, Canada, October 1986.